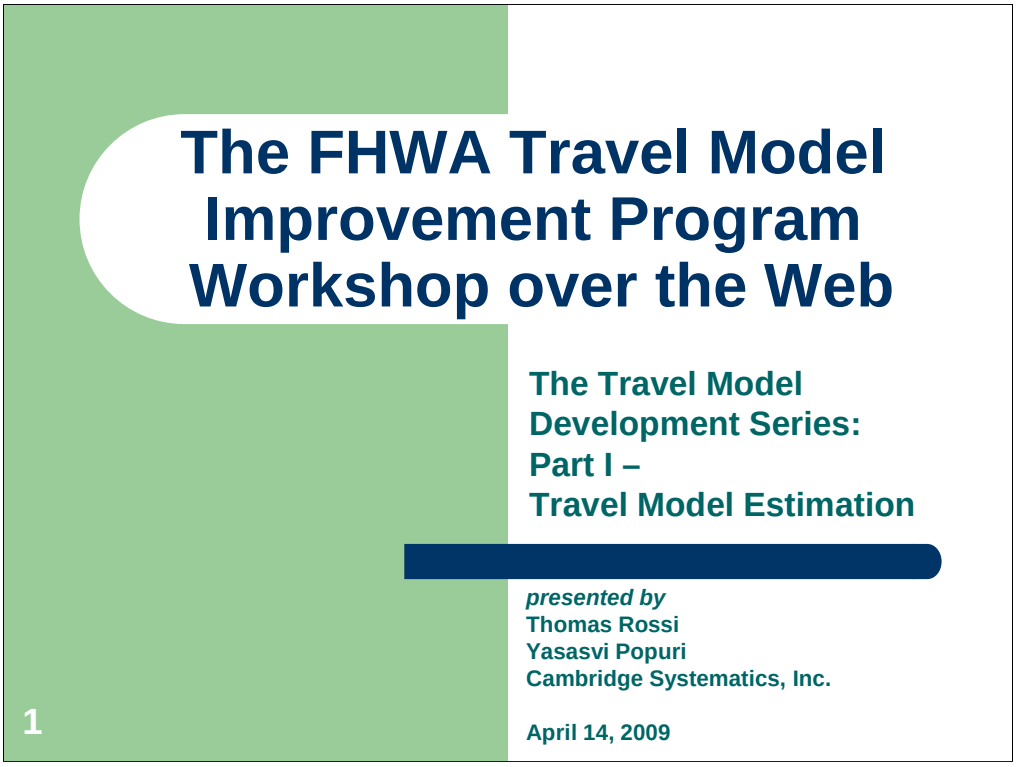
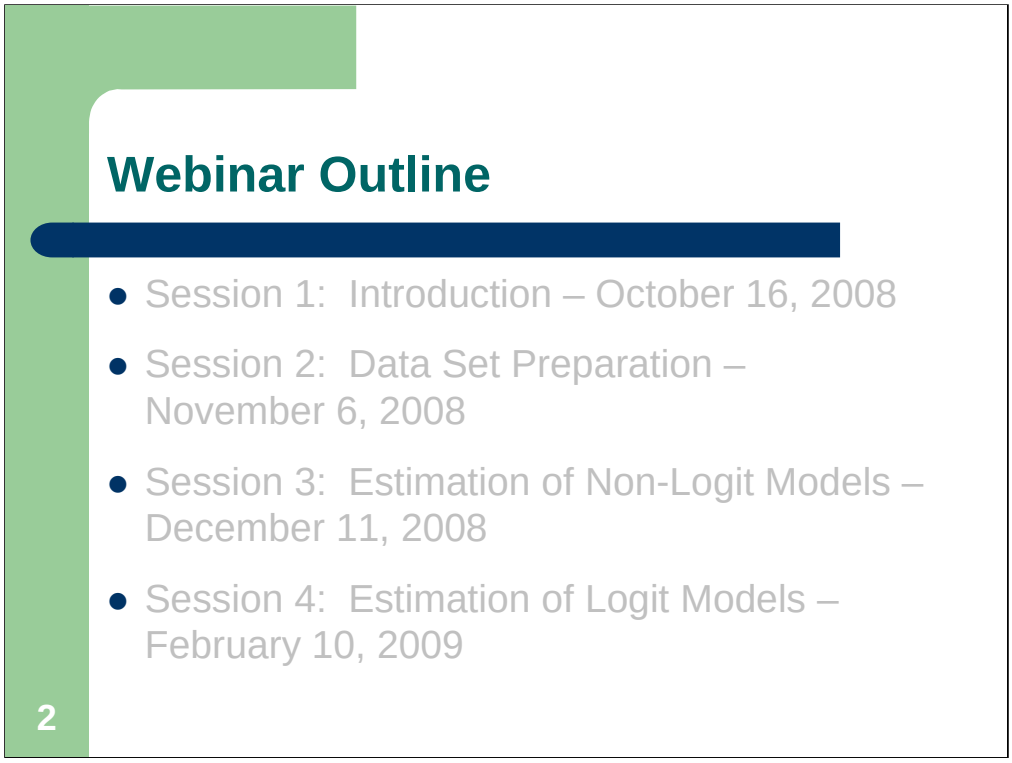


Key Message: Purpose of the Webinar Series

Details:

Welcome to the FHWA TMIP Workshop over the Web. This workshop is targeted at Transportation modelers who have a low to moderate level of familiarity with the estimation and validation of travel models.

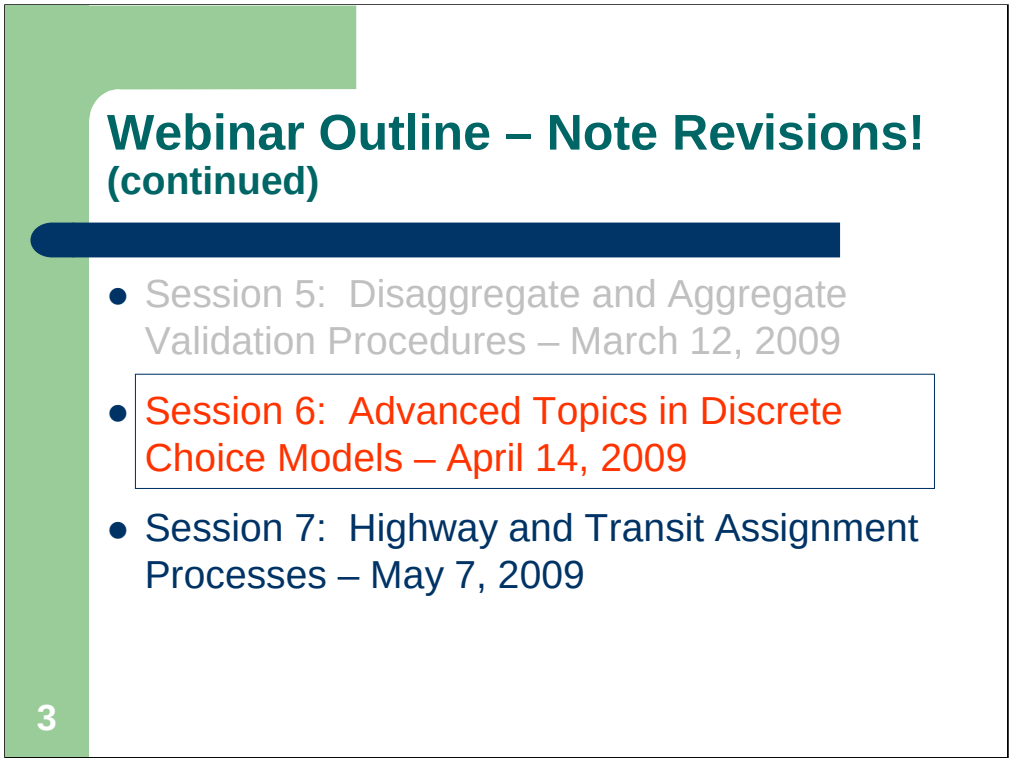
This series of webinars will introduce the development of model estimation data sets, the structures of the various model components, and the procedures for estimating models. The workshop will include lectures, discussion, and "homework," that participants will be expected to complete between sessions.



Key Message: Past Sessions

Details:

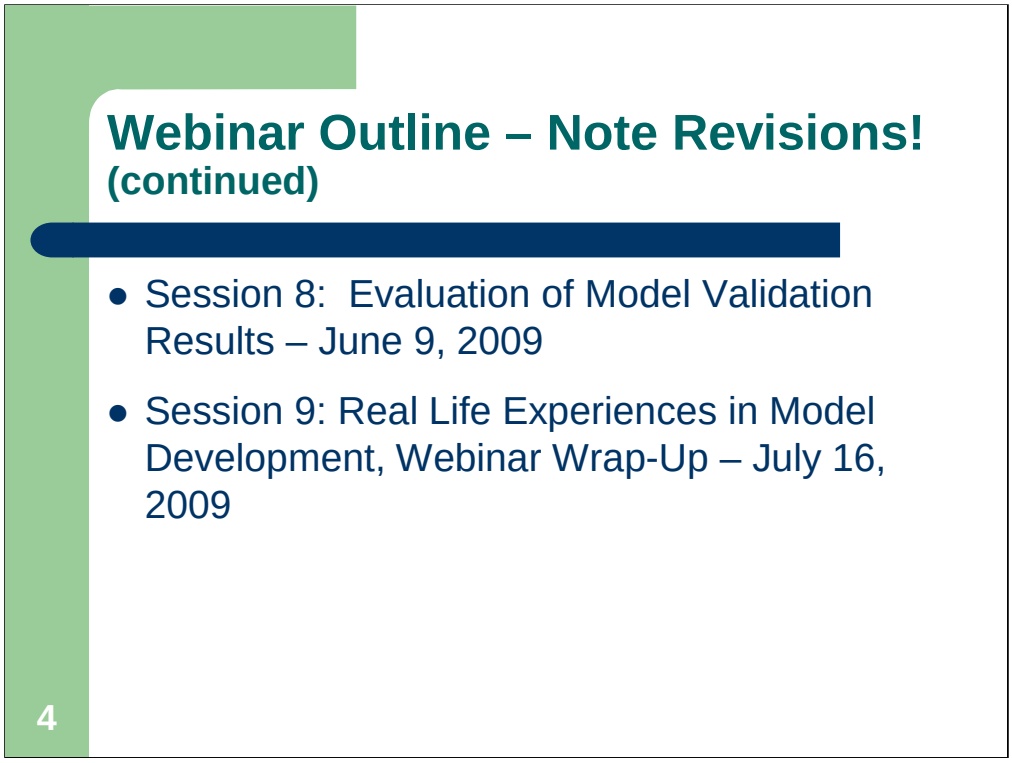
This is the webinar outline. These sessions have already occurred.



Key Message: Upcoming Sessions

Details:

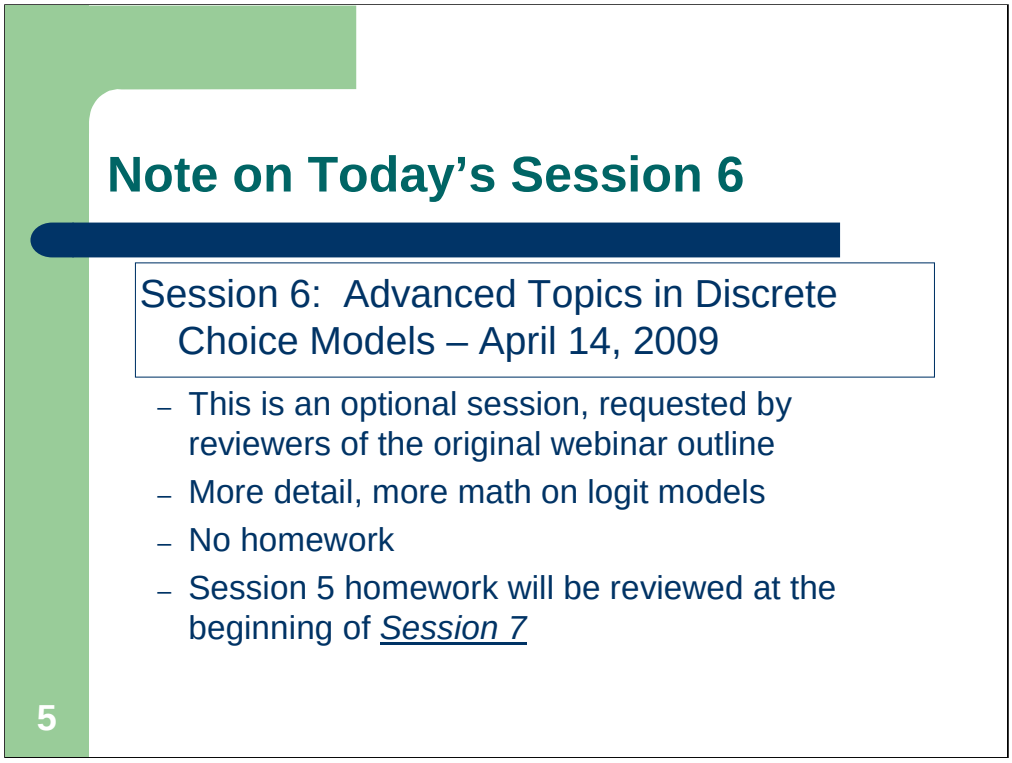
After today's Session 6, Session 7 will be conducted on May 7.



Key Message: Upcoming Sessions

Details:

Session 8 will be conducted on June 9 and a new Session 9, on real life experiences in model development and the webinar wrap-up, will be held on July 16.



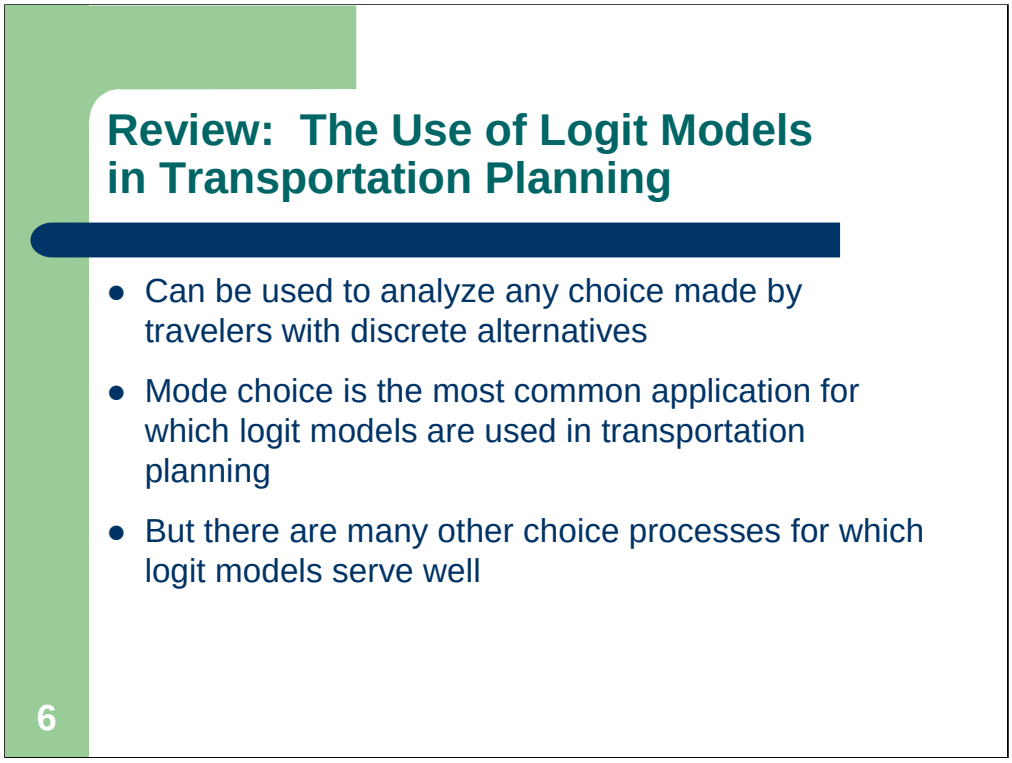
Key Message: Session 6

Details:

A note on today's session. This was requested by reviewers of the original outline for this webinar for people who want to get into more detail on logit models. Topics will include:

- Modeling disaggregate individuals
- More on generic vs. alternative-specific variables
- Interpreting model estimation results
- Examples of advanced variable specification
- Likelihood functions
- Application programming for logit models

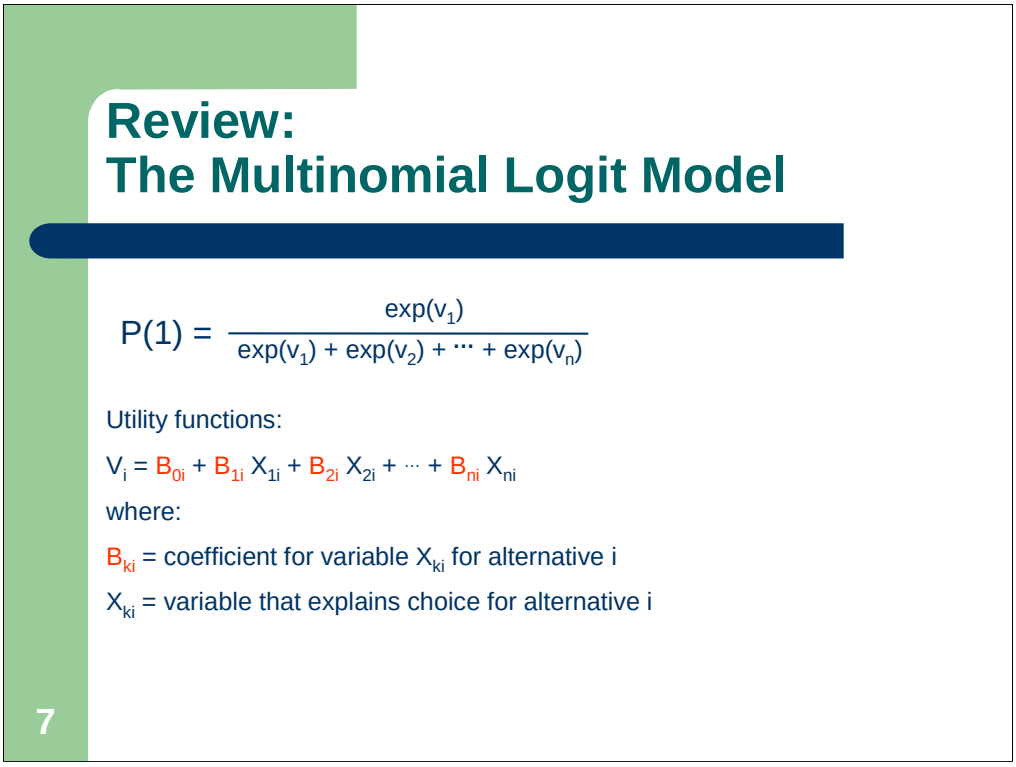
Because this is an optional session, there is no homework associated with it. Therefore, the homework for this session will be reviewed at the beginning of Session 7 in May.



Key Message: Logit Model Uses

Details:

In Session 5 we discussed the use of logit models in travel demand modeling. Logit models are used to model any choice with discrete alternatives (modes, number of vehicles available, destination zones, etc.). While mode choice is by far the most common application for logit models in travel modeling, they can be used in many other types of models.

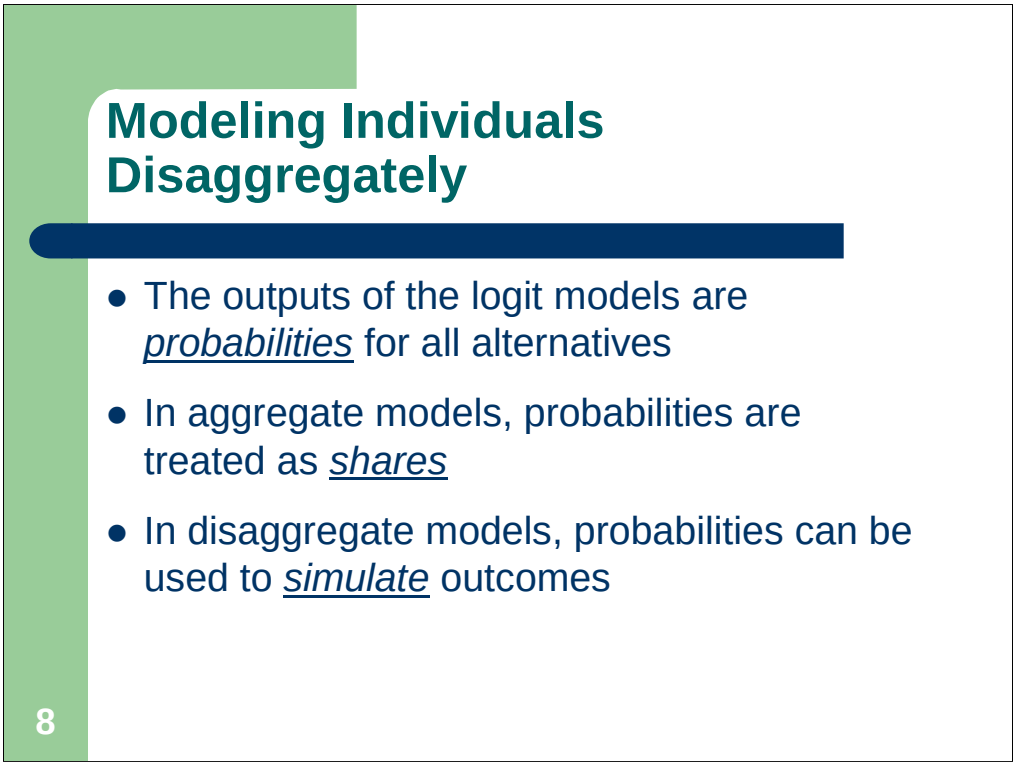


Key Message: The Multinomial Logit Model

Details:

The multinomial logit model estimates the probabilities of choosing among any number of alternatives. The probability of choosing an alternative is a function of the deterministic components of the utilities of all alternatives. The utilities are exponentiated, and the probability of each alternative is the exponentiated utility divided by the sum of the exponentiated utilities for all alternatives. The sum of the probabilities must be one and the probabilities must be between zero and one, and if the utilities are finite numbers, no alternative has a probability of exactly zero or one (although they can be very close).

The utility function used in logit models is a linear combination of the variables that are used to explain the choice being analyzed. The coefficients in the utility function are estimated parameters.

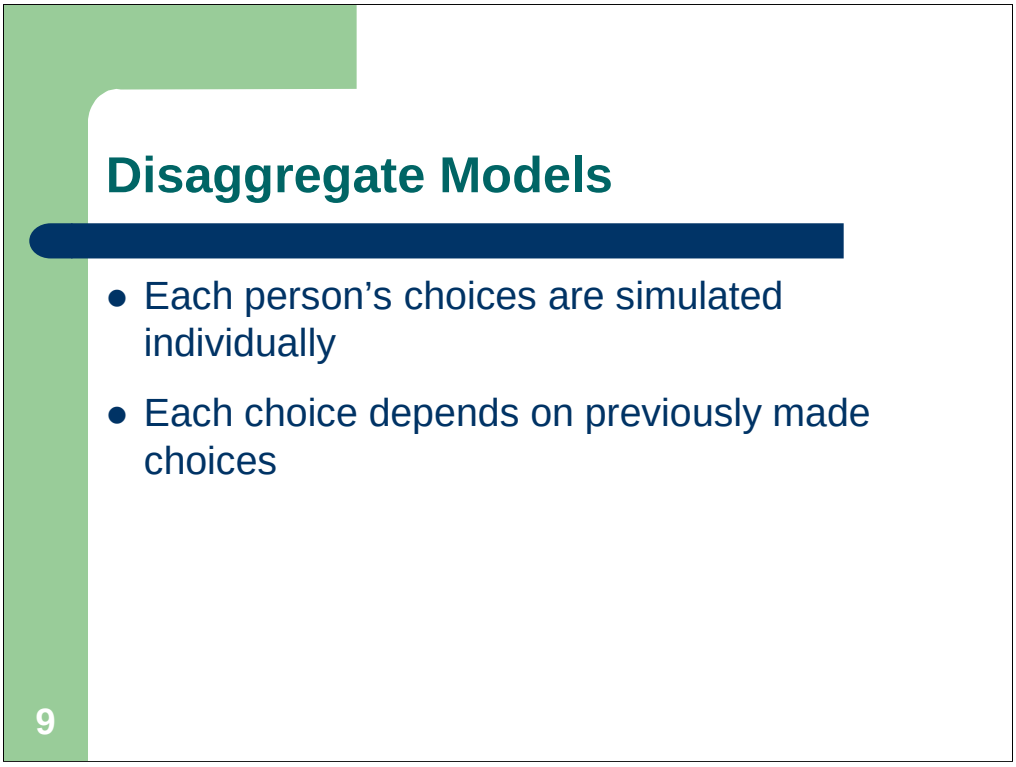


Key Message: Disaggregate Modeling

Details:

As the equation on the last slide shows, the outputs of logit choice models are probabilities for all available alternatives. In aggregate models, such as conventional four-step models, we use these probabilities as shares to apply to the market segment to which the choice probability applies. For example, in a mode choice model, the probability of using transit with walk access would be computed separately for each origin-destination zone interchange for each trip purpose. So if that probability is, say, 10%, then 10% of the trips in the appropriate O-D cell in the trip table for that purpose would be deemed transit with walk access trips. If there is further segmentation in the model, say by income level, then the trip table would be segmented by income level, and the probability computed for each income level would be applied to the trip table cell for that income level.

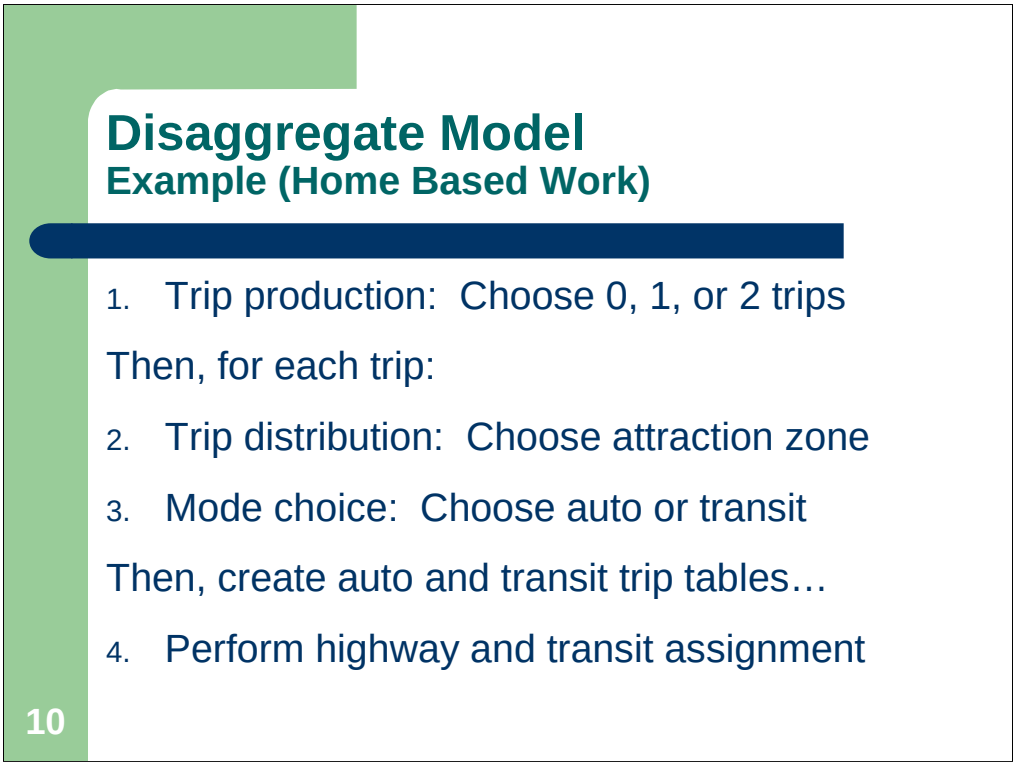
In a disaggregate model, such as a modern activity based model, each person and each tour is simulated individually. So we have a probability that each particular tour or trip uses a particular mode. In this case, we use the probability to simulate an outcome for that particular tour or trip.



Key Message: Disaggregate Models

Details:

So in a disaggregate model, each person's choices are simulated individually, based on the probabilities that come from the logit choice model. Each choice depends on the simulated outcomes of previously modeled choices. For example, a household may be simulated to own one car. All of the subsequent choices, such as mode and destination choice, are simulated based on the household owning one car. The mode choice is simulated based on the previously simulated destination.



Key Message: Disaggregate Model

Details:

The example shown on the next few slides shows how a four-step model might be applied in a disaggregate manner.

For the trip generation model, say there is a logit model where the alternatives are to make zero, one, or two work trips.

For the trip distribution model, applied for each trip that the trip generation model results say occurs, the attraction zone is chosen, given the known home zone of the traveler.

For the mode choice model, a logit model is used to simulate which mode is chosen.

Then, based on the results of previous models, we have a list of trips and their origins, destinations, and modes, and so trip tables for assignment can be created.

Disaggregate Model
Example (continued)

1. Trip production: MNL (3 alts.)

$U_0 = 0$

$U_1 = B_{10} + B_{11}$ (adult) $+ B_{12}$ (worker) $+ B_{13}$ (high inc.) $+ B_{14}$ (med. Inc.) $+ B_{15}$ (male)

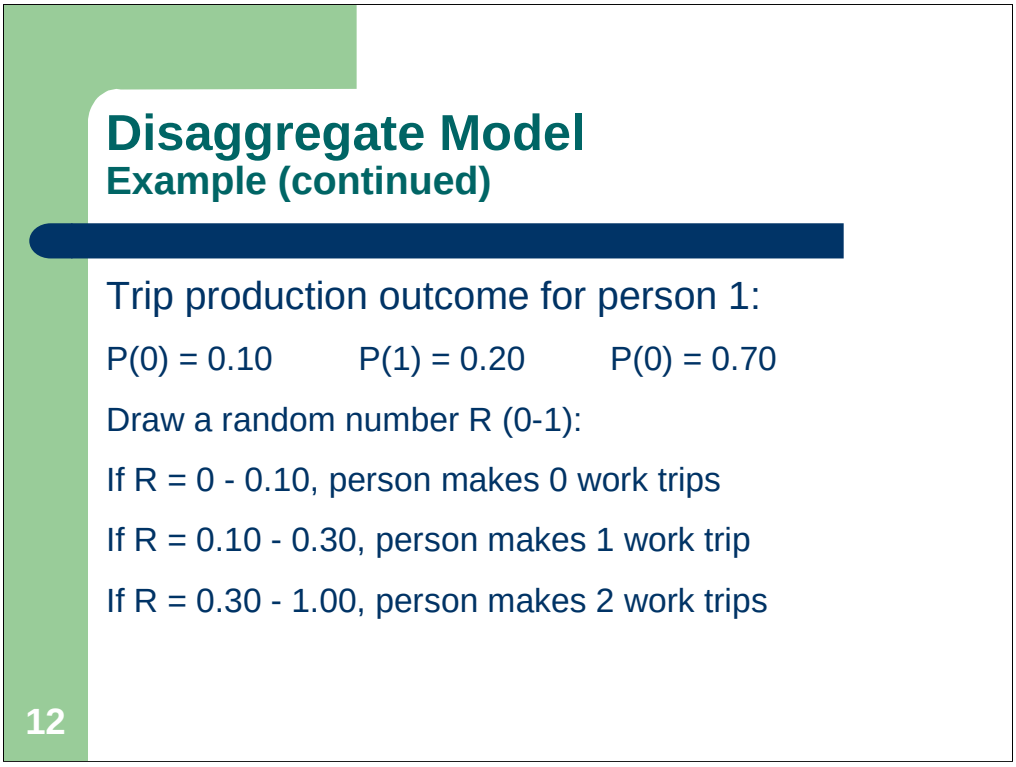
$U_2 = B_{20} + B_{21}$ (adult) $+ B_{22}$ (worker) $+ B_{23}$ (high inc.) $+ B_{24}$ (med. Inc.) $+ B_{25}$ (male)

11

Key Message: Disaggregate Model

Details:

These are example utility functions for the three alternatives for the home based work trip production model. Whether someone makes zero, one, or two trips depends on whether he is an adult, whether he is a worker, his gender, and household income level. We would expect, for example, for B_{11} and B_{21} to be positive, since one is more likely to make work trips if he is an adult.



Key Message: Disaggregate Model

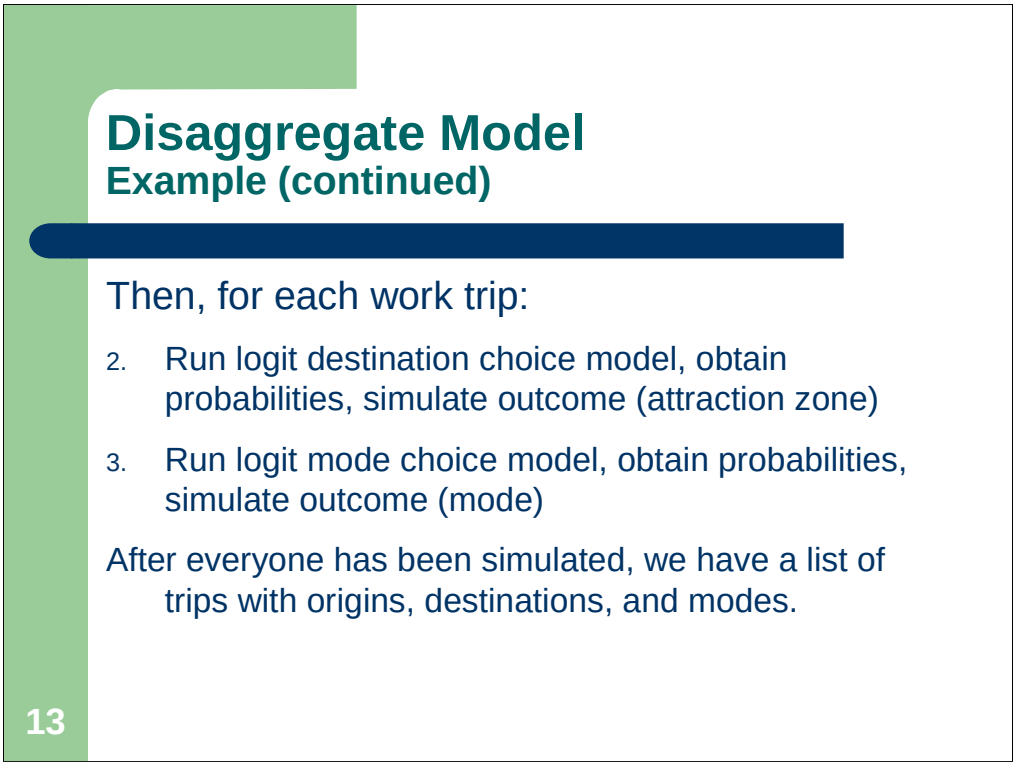
Details:

Let's say that we computed the utilities for an individual and the choice probabilities came out to 0.10 for zero trips, 0.20 for one trip, and 0.70 for two trips. To simulate the number of trips this person makes, we draw a random number between zero and one.

If the random number is between 0 and 0.10, the person makes 0 work trips.

If the random number is between 0.10 and 0.30, the person makes 1 work trip.

If the random number is between 0.30 and 1.00, the person makes 2 work trips.



Key Message: Disaggregate Model

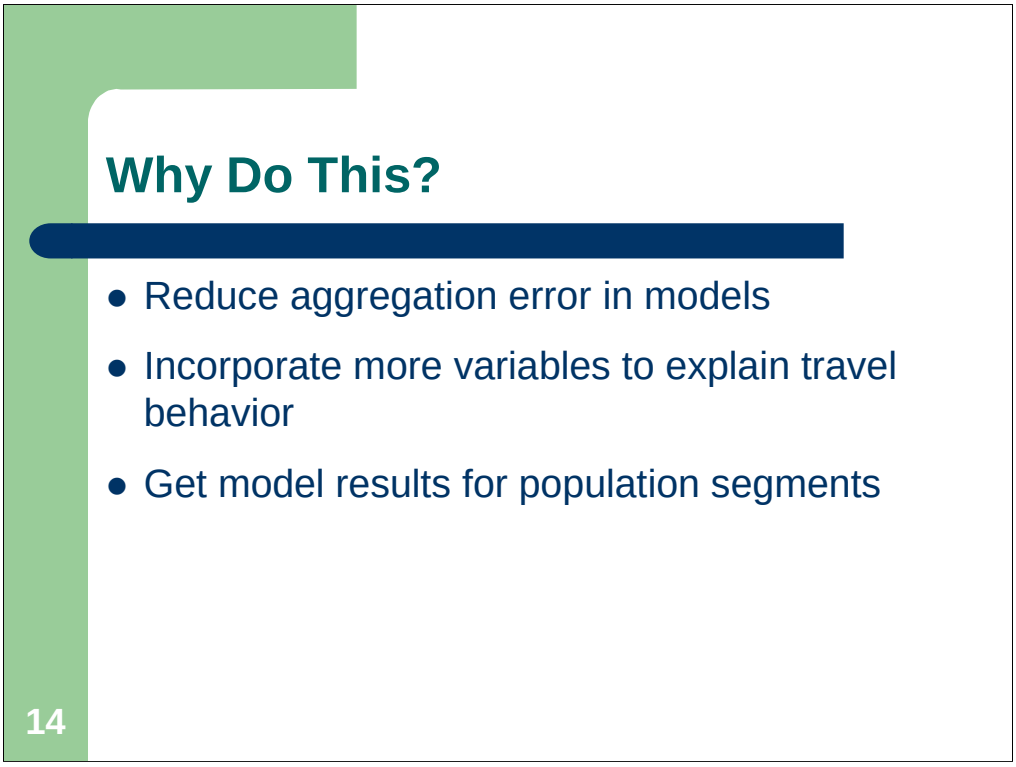
Details:

So now we know the number of work trips this person makes. If it is one trip, we get from the destination choice model the probabilities for each attraction zone to which the person might be traveling. We draw another random number and determine which zone the trip is attracted to. If there are two trips made, we simulate the zone for each.

(This points out one of the problems with trip based models. It would seem very likely that if someone is making two work trips, they are attracted to the same workplace and they represent the round trip between home and work. But unless the model is tour based, it might not guarantee this type of outcome. It would be possible to have the model simulate only one work destination for each traveler and apply it to all work trips. But if there are some travelers who do have different attraction locations—say coming straight home from an off-site meeting—this could not be modeled. The survey data would have to be examined to see whether the constraint of one work destination per traveler is consistent with most behavior.)

For mode choice, for each trip we know the production and attraction zones. We get the probability of using auto and transit for that traveler between those two zones, and we draw another random number to simulate which mode is chosen.

After this is done for every person, we have a list of trips with origins, destinations, and modes. This is enough information to create trip tables for assignment.



Key Message: Disaggregate Model – Why is it a good thing?

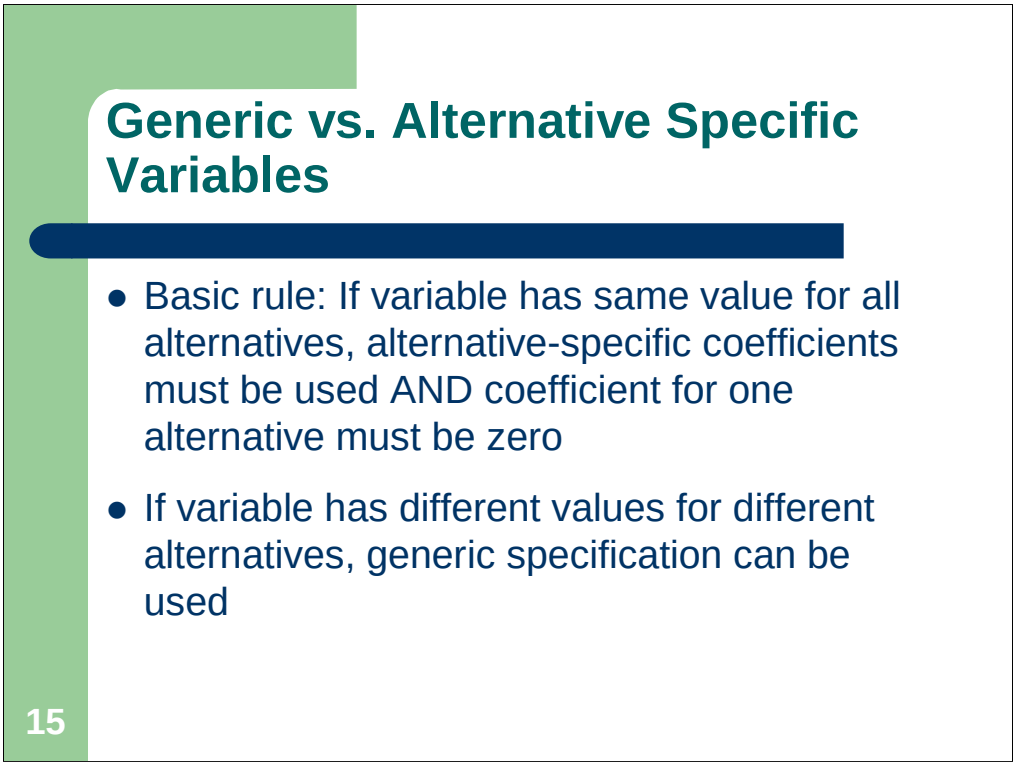
Details:

Why would we bother applying the model in a disaggregate manner?

First, we reduce aggregation error. Anytime you aggregate data you introduce aggregation error. Modelers have been trying to reduce aggregation error for many years, introducing additional market segmentation, smaller zones, etc.

Second, in a disaggregate model, you can have many more explanatory variables. You are limited only by the data. Recall the utility function for the example trip production model we just saw. It included four variables: adult status, worker status, income level, and gender. Typically our trip production models are cross-classification models with two or possibly three variables. Having four variables would be impractical. Our example could have included even more variables, any for which survey data were collected and forecasts are available.

Another benefit is the ability to get results for any population segment for which we have data on the segmentation variable. For example, if we want to see the effect of a policy alternative on low-income households, we can sort the list of results by the income level of the traveler's household. This can be valuable in analyses such as environmental justice.

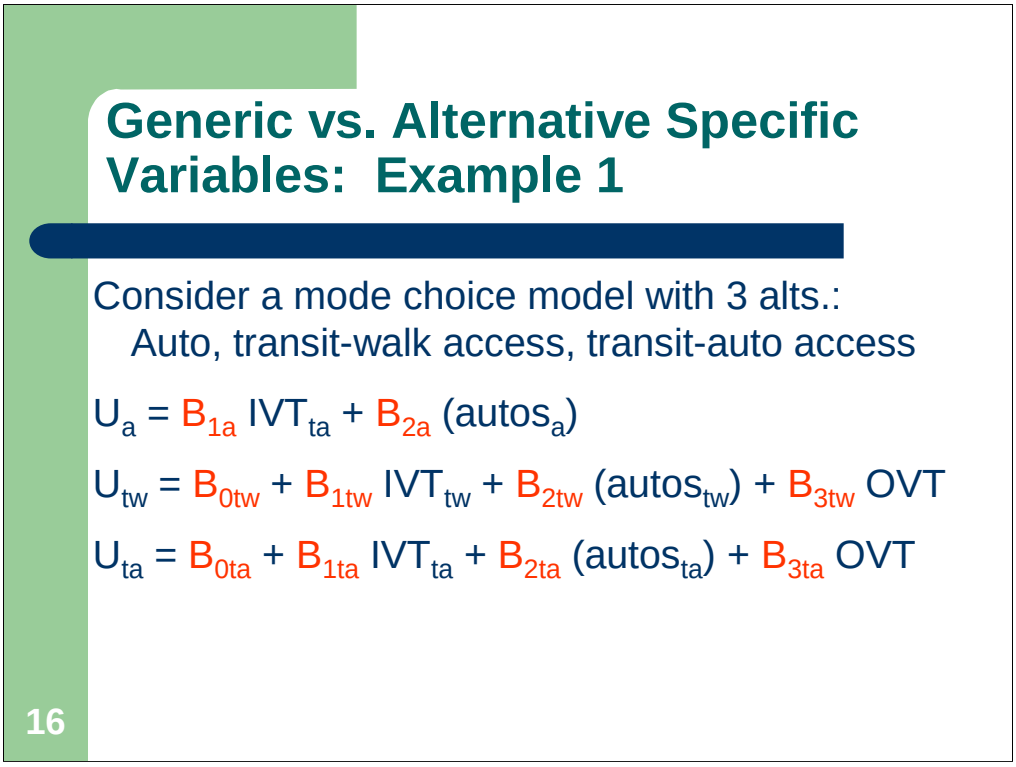


Key Message: Variable Types and Specifications

Details:

We have discussed including in utility functions generic variables, where the variable has the same coefficient in all utility functions (or for more than one), and alternative specific variables, which have different coefficients for all alternatives. When can or should we use generic variables, as opposed to alternative specific?

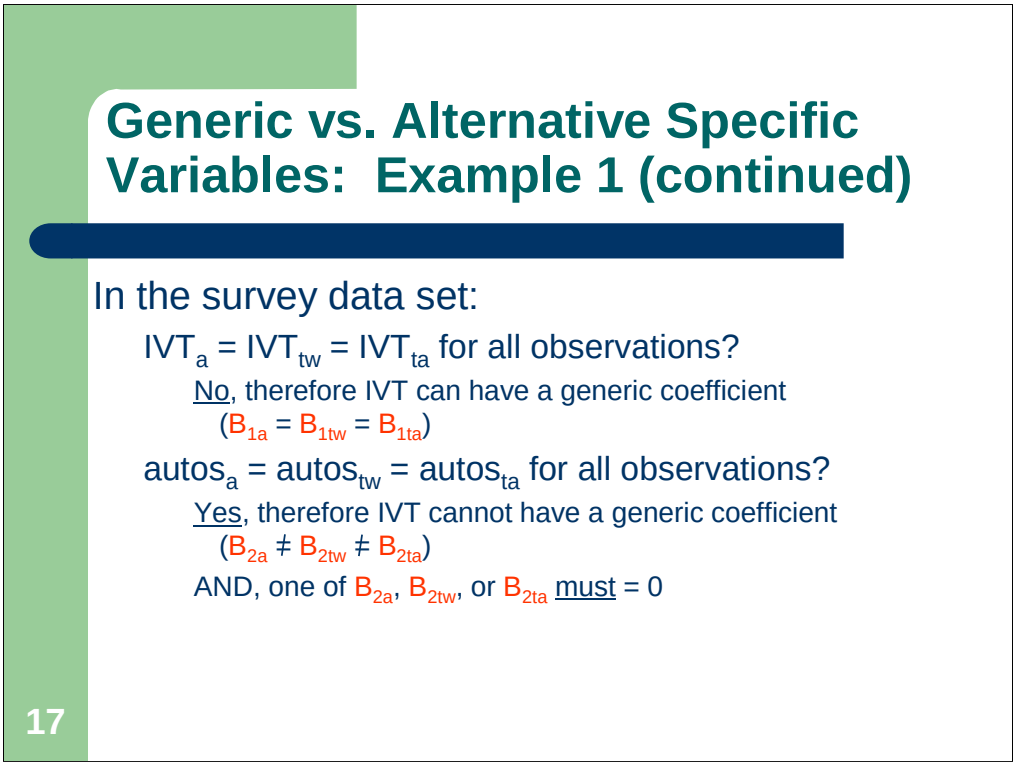
The basic rule is: If a variable has same value for all alternatives, alternative-specific coefficients must be used for this variable AND the coefficient for one of the alternatives must be zero. If a variable has different values for different alternatives, a generic specification can be used. Let's look at an example.



Key Message: Variable types and Specifications - Example

Details:

Consider the mode choice model shown here with three alternatives: auto, transit-walk access, and transit-auto access. There are three variables, in-vehicle time, number of autos, and out-of-vehicle time, along with alternative specific constants. As we already know, one of the three alternatives must be chosen to have a constant of zero. We have chosen the auto alternative.

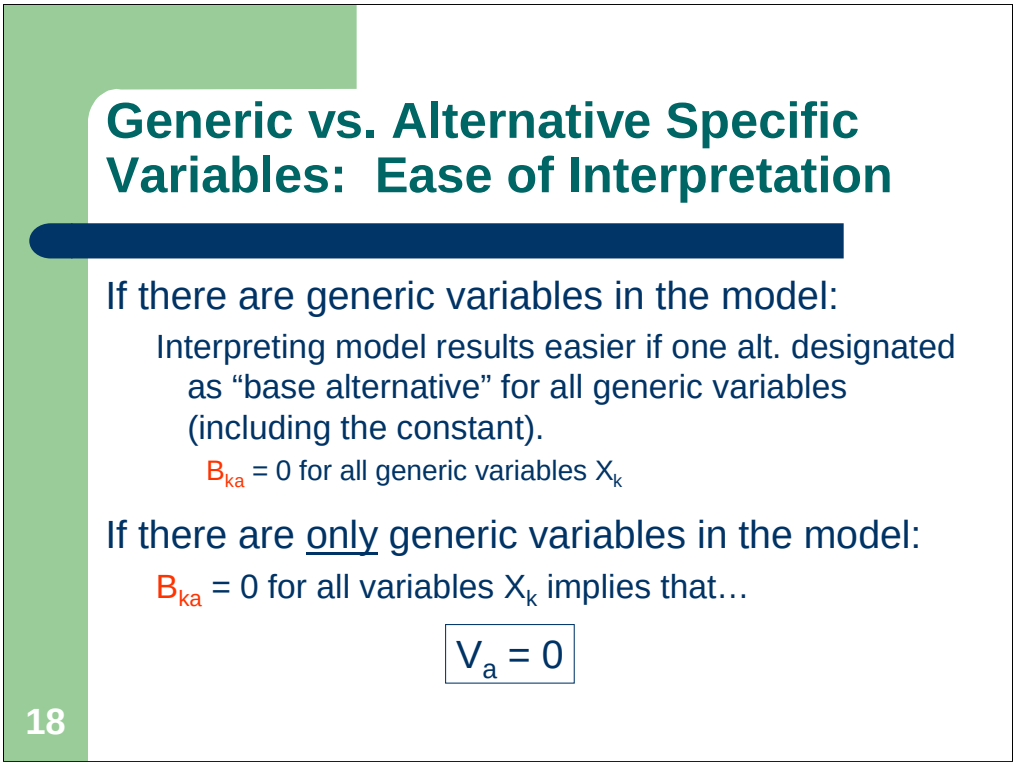


Key Message: Variable types and Specifications - Example

Details:

In the survey data set, does the in-vehicle time have the same value for each modal alternative? No, the auto time is usually not equal to the transit time, and the transit-walk access time is not equal to the transit-auto access time. They are, in effect, three separate variables. So we can make the IVT coefficient the same for all three alternatives. We can estimate the model with different IVT coefficients if we choose, but we don't have to.

Now, does the number of autos variable have the same value for each modal alternative? Yes, each household owns the number of autos reported in the survey data regardless of which mode was chosen. So we cannot make the autos coefficient the same for all three alternatives. We must use three different values for the autos coefficient, and one of them must be zero.



Key Message: Variable types and Specifications - Interpretation

Details:

If there are generic variables in the model, interpretation of model results is easier if one alternative is designated as the "base alternative" for all generic variables. We include the constant in this set of generic variables because it can be interpreted as a variable that has the value 1.0 for all observations in the data set multiplied by its coefficient (the estimated value of the constant). Therefore:

B_{ka} = 0 for all generic variables X_k

If there are only generic variables in the model, the coefficient for every variable in the base alternative's utility will equal zero (including the constant). Therefore the utility of that alternative is always zero.

Vehicle Availability Model
Alternative Specific Variables

	Vehicle Availability Level				
Variable	0	1	2	3	4+
Persons per household	--	--	0.1164 (2.1)	0.1164 (2.1)	0.2571 (2.1)
Workers per household	--	--	0.4915 (5.2)	1.474 (10.8)	2.139 (10.0)
Household density	--	-0.0458 (-2.9)	-0.1327 (-5.4)	-0.1717 (-4.4)	-0.2549 (-3.0)
ln(income)	--	1.130 (8.7)	2.497 (13.9)	2.995 (12.7)	3.242 (7.6)
Transit/highway accessibility	--	-1.133 (-1.7)	-2.054 (-2.8)	-2.742 (-3.3)	-2.742 (-3.3)
Persons less than vehicles	--	--	-2.870 (-8.8)	-1.017 (-5.3)	-0.5181 (1.1)
Constant	--	0.164 (0.2)	-3.761 (-4.6)	-8.229 (-8.0)	-12.87 (6.8)

ρ^2 w.r.t zero = 0.447 ρ^2 w.r.t constants = 0.302

19

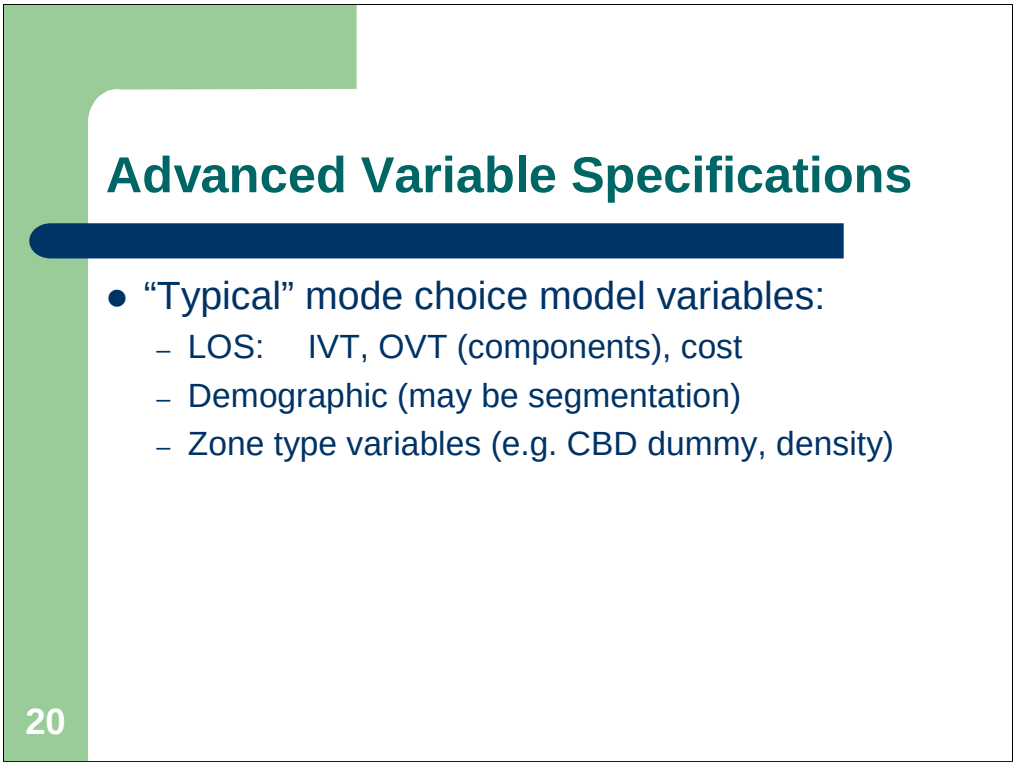
Key Message: Vehicle Availability Model – Alternative Specific Variables

Details:

In Session 4, we saw a model with only generic variables and a base alternative with utility zero. In this vehicle availability model, there are five alternatives corresponding to the household owning zero, one, two, three, or four or more vehicles. All of the variables are generic. The household characteristics are the reported values from the household survey and do not vary among alternatives. The accessibility and density variables are derived from the network skims and the zonal socioeconomic data and does not vary no matter how many vehicles the household chooses to own.

In this model, the base alternative was chosen to be zero vehicles. All of the estimated coefficients should be interpreted as relative to owning zero vehicles. For example, the increasingly positive coefficients for workers per household means that the utility is higher for increasing numbers of vehicles as the number of workers increases.

In terms of model application results, it doesn't matter which alternative is chosen as the base. If 4+ vehicles was chosen as the base, its utility would be zero, and the coefficients of the zero-vehicle alternative would be -1 times the values shown in this table for the 4+ alternative. The coefficients for each other alternative would be the differences between the value shown for that alternative and the 4+ alternative. For example, the persons per household coefficient for 3 vehicles would be (0.1164 – 0.2571) = -0.1407. Applying that model would yield the same results as the model shown here, since only relative utilities matter.



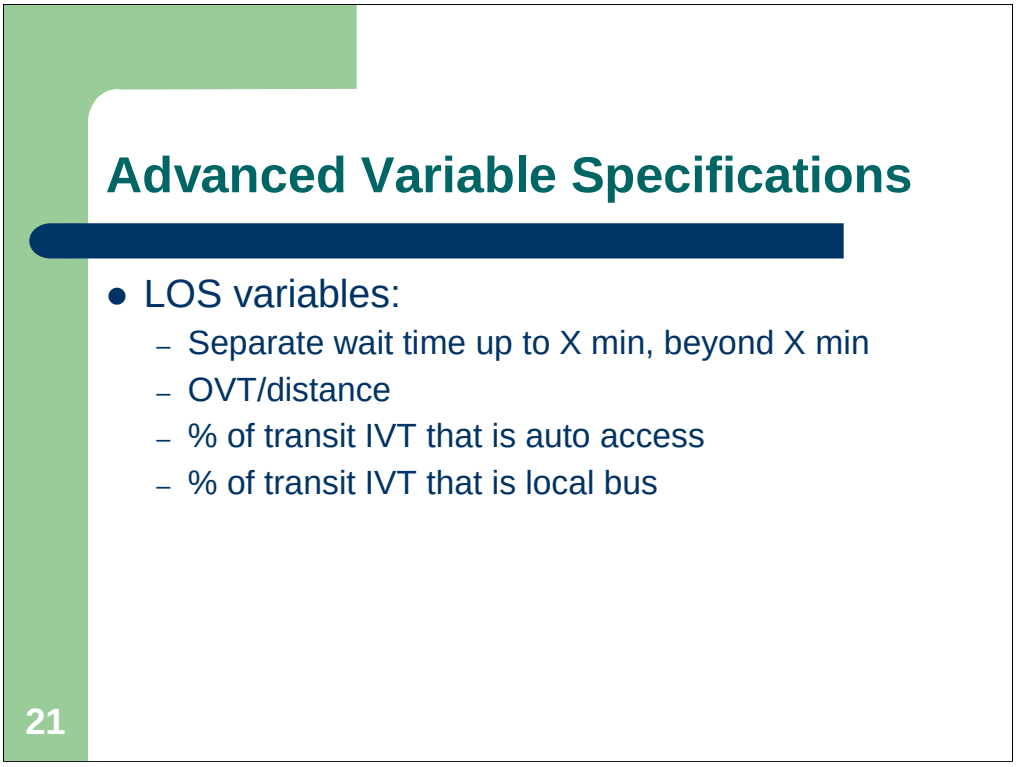
Key Message: Vehicle Availability Model – Advanced Variable Specifications

Details:

Traditionally, mode choice models in the U.S. have used similar variables. They include:

- Level of service variables, such as IVT, OVT (or its components such as wait time, walk time, etc.), and cost (fare, parking, etc.).
- Demographic variables, such as income or auto ownership. These may be segmentation variables (e.g. 1 if zero autos, 0 otherwise).
- Zone type variables, such as CBD dummy, population or employment density, etc.

We can also consider more "advanced" model specifications.

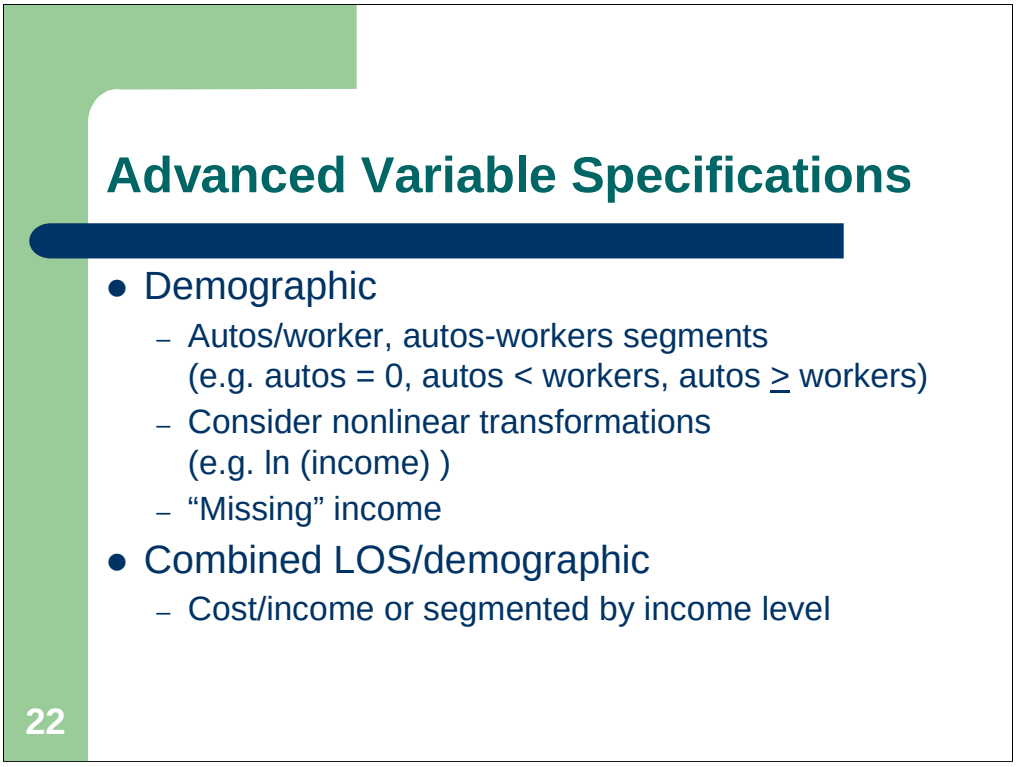


Key Message: Vehicle Availability Model – Advanced Variable Specification Examples

Details:

Some examples of more advanced level of service variable specifications include:

- Separate wait time up to and beyond a specific level, for example 10 minutes. This takes into consideration that wait time is not linear with respect to transit headway; when the headways get large, riders do not arrive randomly at stops.
- Out-of-vehicle time can be divided by distance, to account for the fact that travelers might not find small amounts of out-of-vehicle time as onerous when it is part of a longer trip.
- The % of transit IVT that is auto access can be used to distinguish between transit-auto access trips that are mostly auto or mostly transit.
- The % of transit IVT that is local bus can be used as a variable in multimode transit trips (e.g. bus-rail). Presumably, a trip that has a higher percentage of premium transit time as opposed to local bus is more desirable.



Key Message: Advanced Variable Specification Details – Demographic and LOS/Demographic

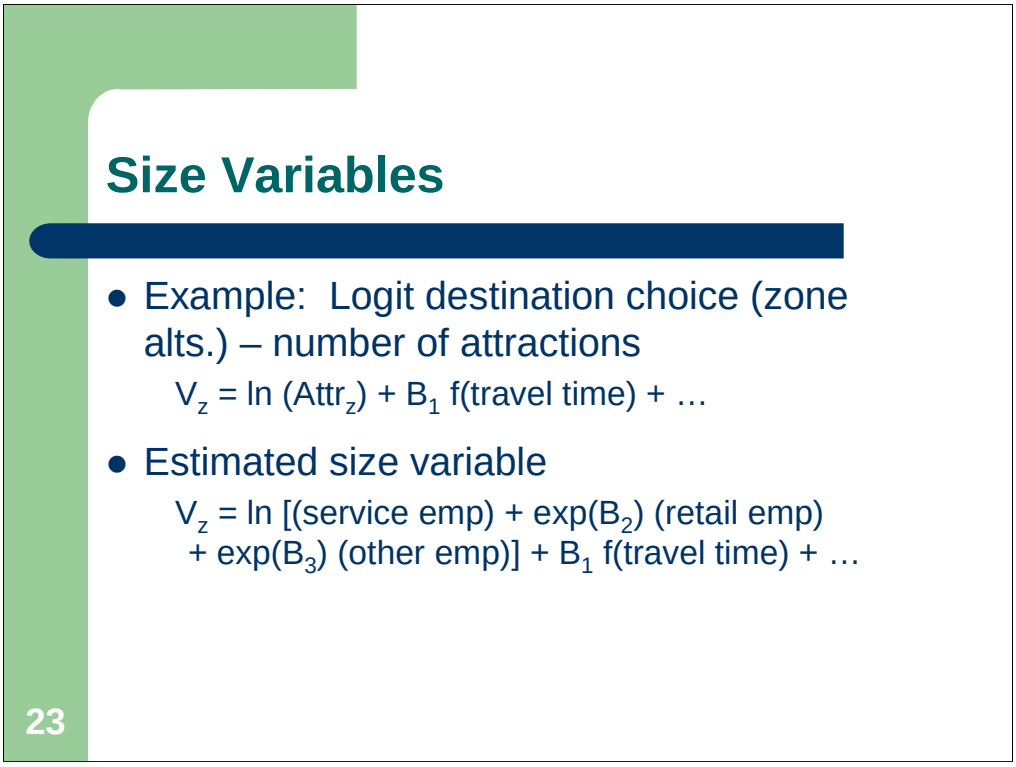
For demographic variables, one can consider:

Autos per worker, or segmentations of the number of autos compared to the number of workers. For example, three segments: autos = 0, autos < workers, autos > workers.

One can consider nonlinear transformations of variables if it is felt that the relationship between different levels of a variable and travel choice is not linear. For example, if it is felt that the difference between $25,000 and $50,000 in household income has a different effect on mode choice than the difference between $25,000 and $50,000, one could try the natural log of income, or some other transformation.

Usually a fairly significant percentage of respondents decline to report their incomes. One way of using the other information in these observations is to define a "missing" income variable, where the value is 1 if the income was not reported and 0 if it was reported. This is most effective if other income segments are used.

One can also combine level of service and demographic variables. For example, a common variable is cost divided by income, to capture the effect of increasing value of time as income increases. The cost variable could also be segmented by income level (e.g. cost-low income, cost-medium income, etc.).



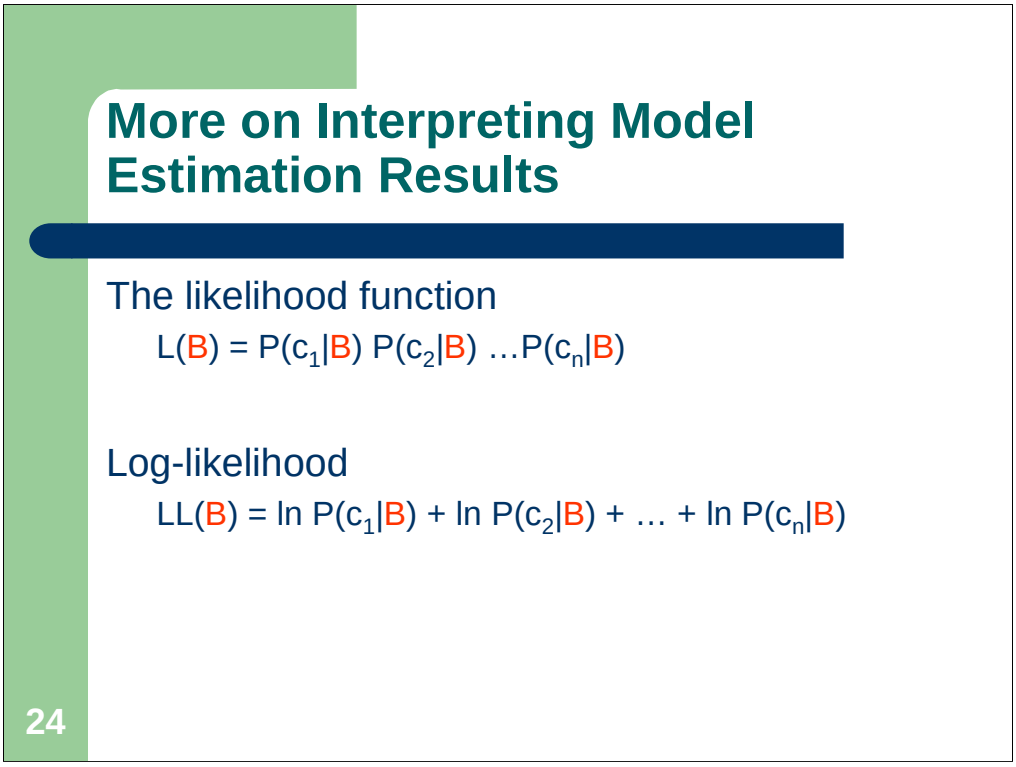
Key Message: Advanced Variable Specification: Size Variables

Details:

Size variables are used in certain types of models such as destination choice models, where the alternatives are actually aggregations of smaller alternatives. A zone can include many destinations for trips. Basically, since zone boundaries are somewhat arbitrary, we don't want the choice of zone boundaries to affect the estimation and application of the logit model.

The size variable is a function of measurable attributes of the aggregate alternative. A common one is the number of attractions in a zone, if there is a good trip attraction model. In this case, the variable is used, in logarithmic form, as shown in the equation.

The size variable can also be defined in the model estimation process. For example, the size variable could be a linear function of service employment, retail employment, and other employment. The form shown here allows the model to estimate the relative contributions of the components of the size variable (in this case, the employment by type) to the "attractiveness" of a destination zone.



Key Message: Interpretation of Model Estimation Results

Details:

In Session 4, we did not have time to go over the details of the likelihood function in logit models.

The likelihood function is the joint probability that the observed choices in the estimation data set were made. Since the probability of each choice is assumed to be independent of the others, the likelihood function is the product of the probabilities of the individual choices.

In the logit function, the probability of each choice is a function of the vector of coefficients in the utility function, B. The logit estimation program chooses the values of B that maximizes the value of the likelihood function.

Because it is easier computationally and it is equivalent mathematically, the program maximizes the logarithm of the likelihood function, or "log-likelihood." Since the probabilities are all between zero and one, the log-likelihood must be less than or equal to zero.

Likelihood Function
Example

Consider a binary logit model, auto vs. bus

Let $V_m = a\ (IVT_m)$

Consider a 3 trip sample:

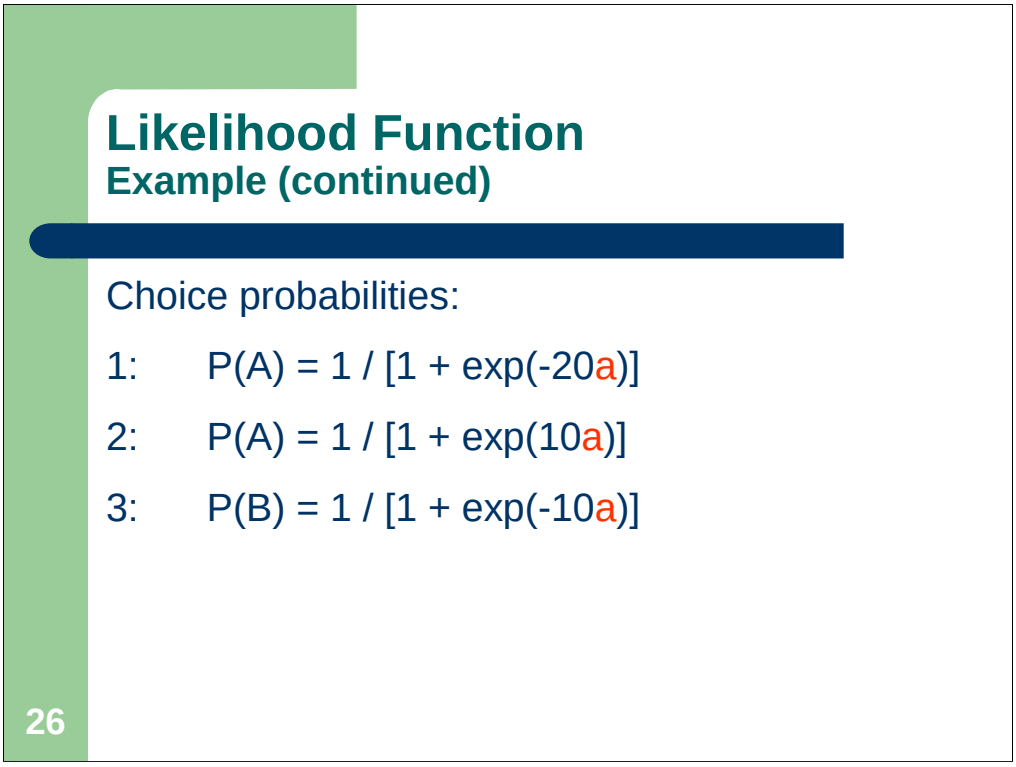
Trip	Choice	IVT_a	IVT_B
1	A	50	30
2	A	10	20
3	B	30	40

25

Key Message: Likelihood Function

Details:

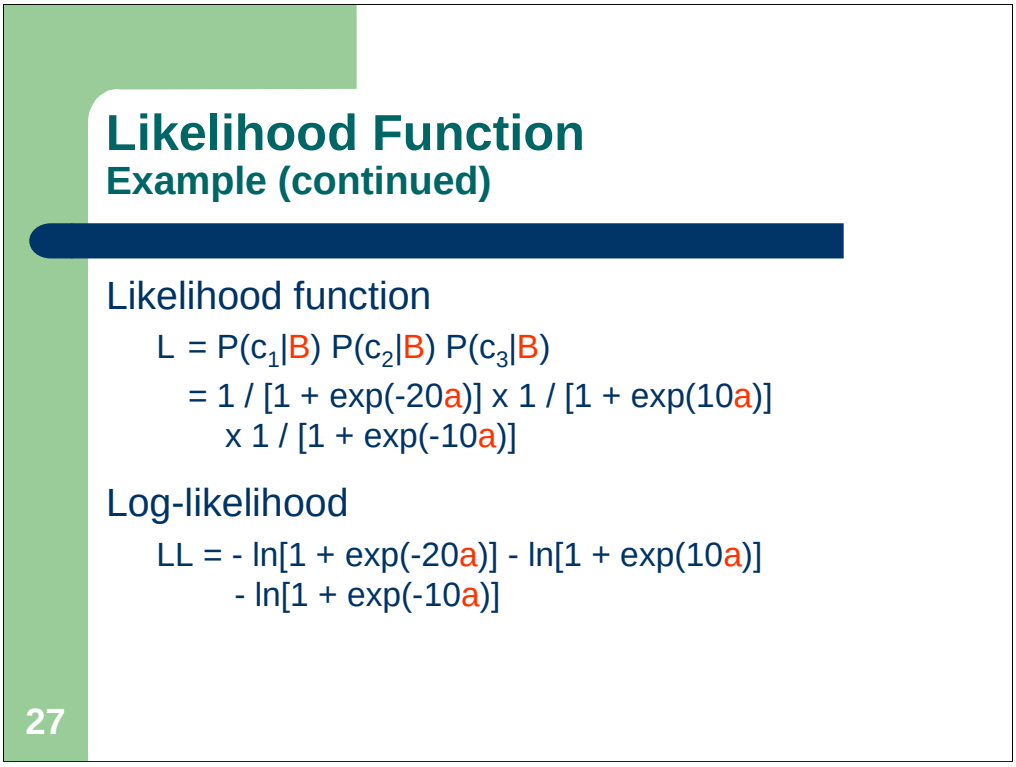
Here is a simple example. Let's say we have a binary mode choice model, and the utility function includes only one variable, in-vehicle time, with parameter "a." The model estimation data set with three observations with the data shown.



Key Message: Likelihood Function

Details:

Using the logit formula, we can compute the probability of each choice as a function of the parameter a. (Remember, in a binary model, the probability can be expressed as a function of the difference between the two utilities. In this case, the difference is the exponent of the differences between the two in-vehicle times, multiplied by a.)

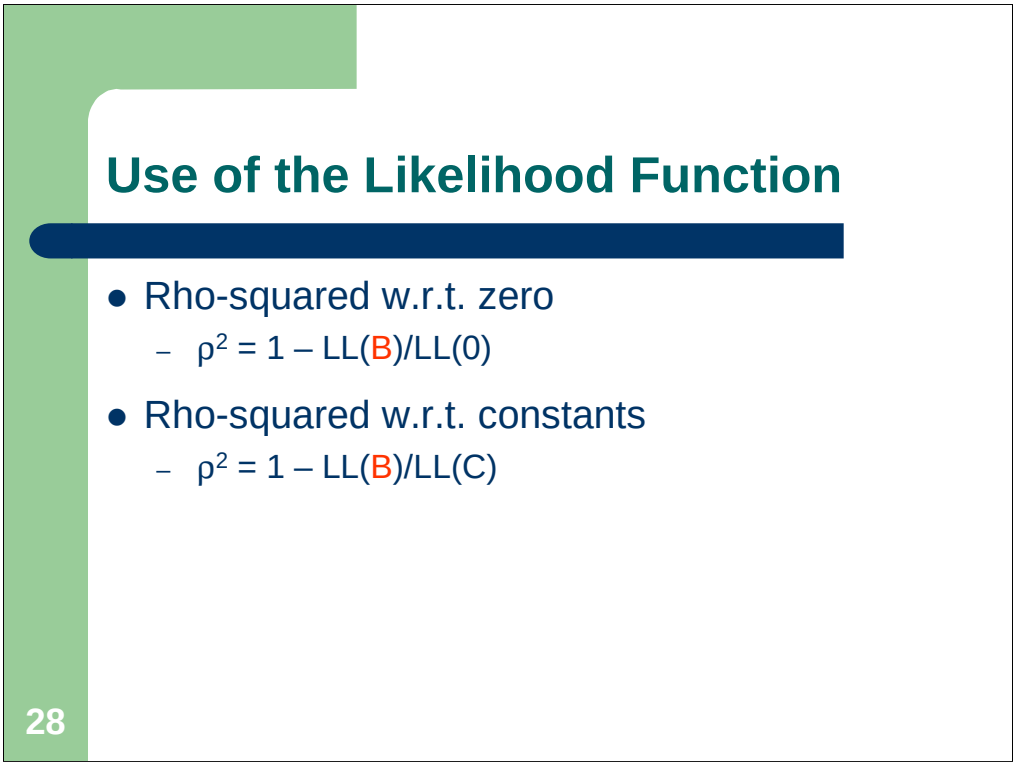


Key Message: Likelihood Function

Details:

The likelihood function can be computed from the formula presented previously, as shown here. The log-likelihood is computed as shown. As can be seen, this is a function of the parameter to be estimated (a, in this case). The optimization process estimates the value of a that maximizes this function. The value of this function is what is shown as the log-likelihood value in model estimation software.

Of course, the likelihood functions are more complex in realistic cases, not only because there are many more than 3 observations, but also because there are several estimated parameters, and the optimization is therefore done for the vector or parameters (B1, B2, etc.).



Key Message: Likelihood Function

Details:

We have seen the "rho-squared" statistic in Session 5, but its derivation was not explained. Here we can see the formulas by which rho-squared is created from the values of the likelihood function at various stages of model estimation.

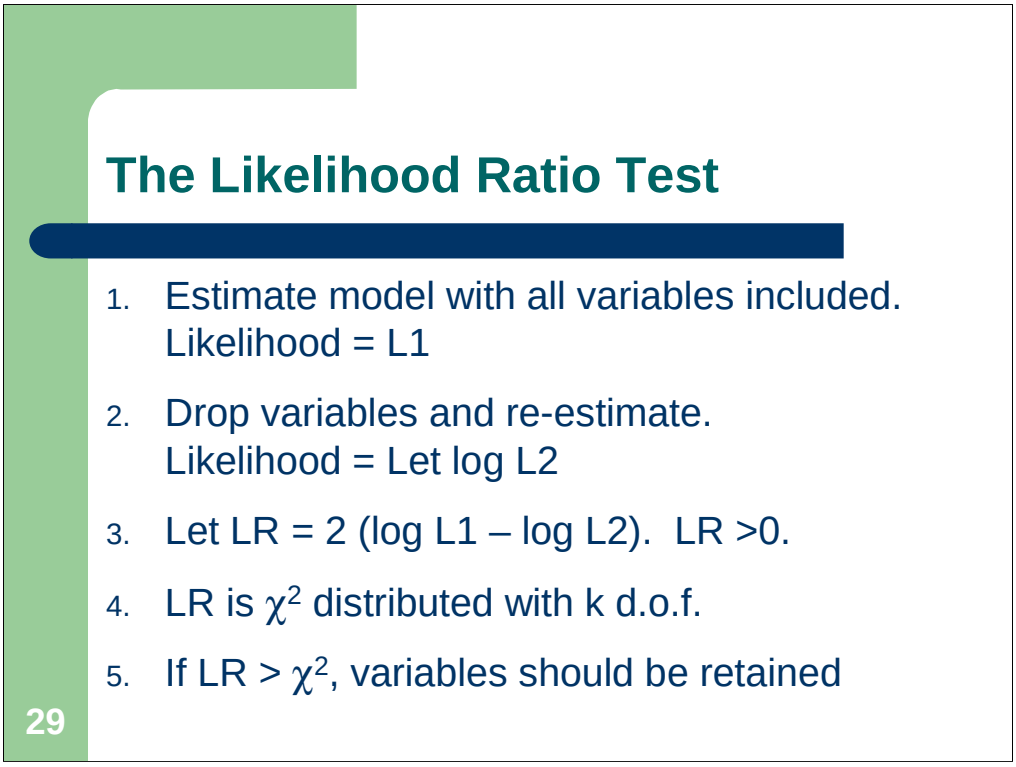
The model estimation software computes the value of the log-likelihood function for the final (maximum likelihood) estimated values for the vector of parameters $B=(B_1, B_2, \ldots, B_n)$. This value, LL(B), is compared to the values of the likelihood function for two "naïve" models. The first naïve model assumes equal probabilities for all alternatives, as if $V_k = 0$ for all alternatives K. This is probably the most naïve model that could be developed, and the value of the log-likelihood function for this model is designated LL(0). We compute "rho-squared with respect to zero" as:

$$\rho^2 = 1 - LL(B)/LL(0)$$

As the formula shows, ρ^2 must be between zero and one since LL(B) > LL(0). The better the model compared to the naïve model, the higher the ρ^2 value.

Another not quite as naïve model is to assume that the probability of each alternative is equal to its share in the data set from which the model is estimated. This is equivalent to a model with only constants, $V_k = B_{0k}$ for all alternatives K. We designate the value of the log-likelihood function for this model LL(C) and compute "rho-squared w.r.t. constants" as:

$$\rho^2 = 1 - LL(B)/LL(C)$$



Key Message: Likelihood Ratio Test

Details:

Besides looking at the statistical levels of significance for individual variables (such as through the use of t-statistics), it is worthwhile to examine whether variables or groups of variables contribute significantly to the explanatory value of the model. The likelihood ratio test provides a way to do this.

The logit estimation software reports the value that the sample log likelihood has when the values of the coefficients equal the maximum likelihood estimates. If the group of variables in question has little explanatory power, then dropping them from the model should have little effect on the maximum value of the log likelihood. Dropping one or more variables always will cause the maximum value of the log likelihood to decrease, but it will not decrease by much if the variables that have been dropped have little explanatory power.

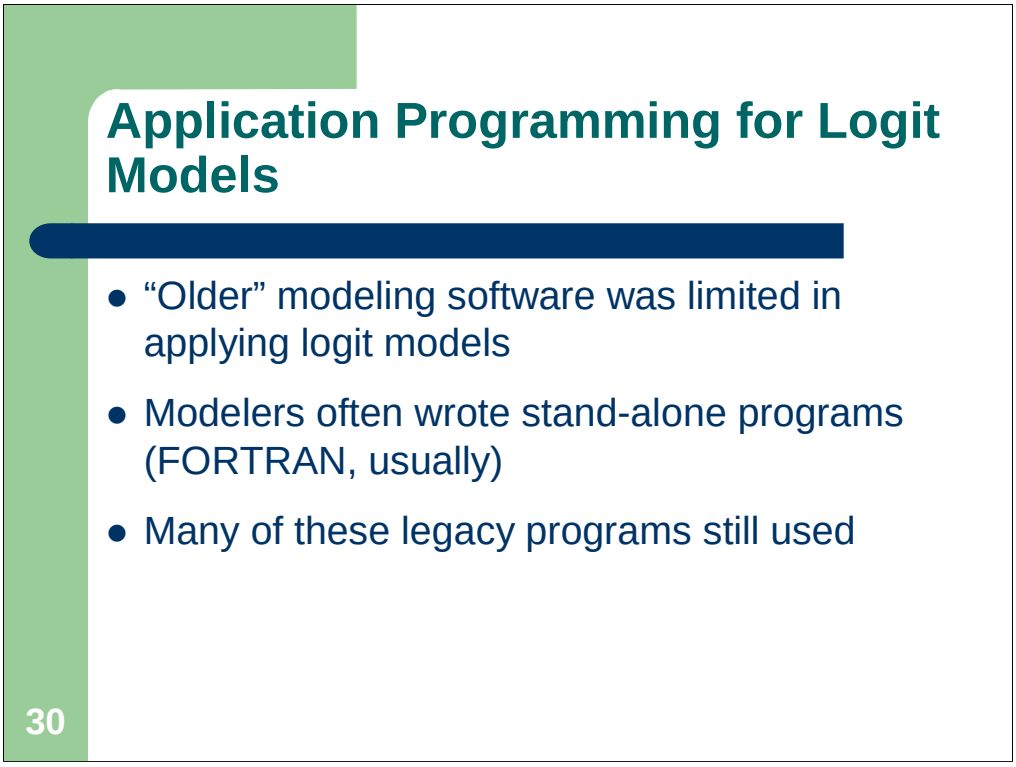
The likelihood ratio test is carried out quantitatively as follows:

1. Estimate the model with all variables included. Let log L1 denote the resulting maximum value of the log likelihood.

2. Drop the variables in question and re-estimate the model. Let log L2 denote the resulting maximum value of the log likelihood.

3. Compute the quantity LR = 2 (log L1 – log L2). LR is called the likelihood ratio test statistic. LR is always a positive number.

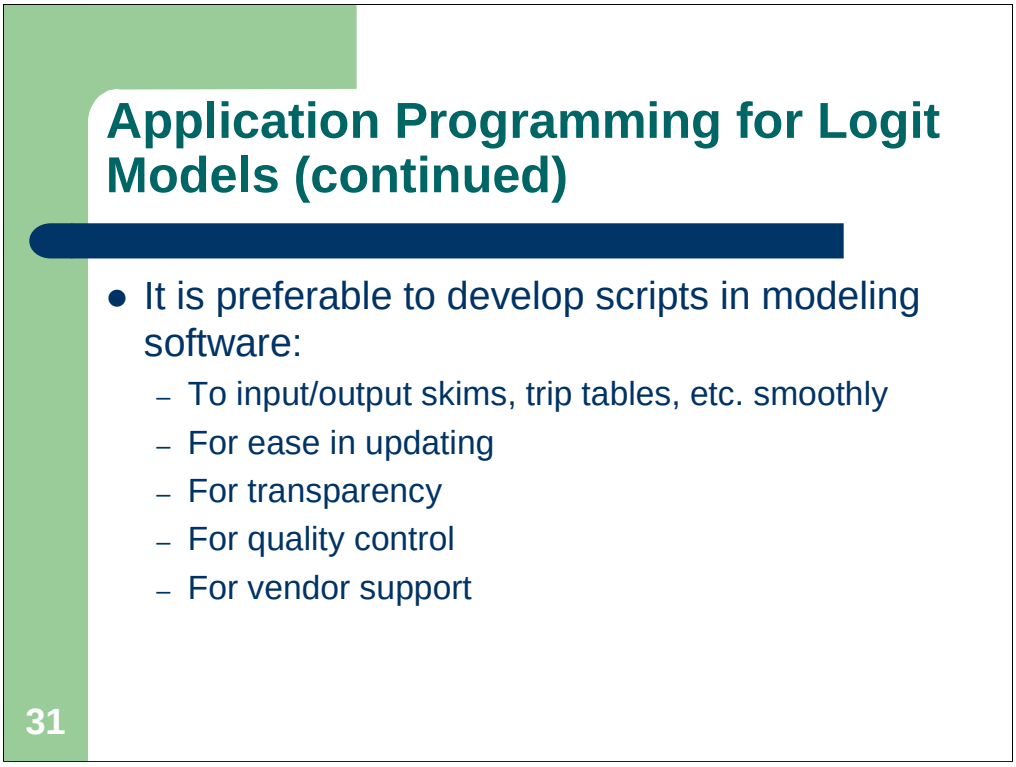
The statistic LR is χ^2 distributed with k degrees of freedom (the number of estimated parameters). If LR exceeds the appropriate critical value from the χ^2 table, then the variables being tested should be retained in the model, even if all of their coefficients have t statistics in the range –1.0 to 1.0. If LR is less than the critical value, then it may be desirable to drop the variables from the model.



Key Message: Programming for Logit Models

Details:

We now discuss application programming for logit models. Older modeling software had limitations in applying logit models, and so modelers often had to write their own programs to apply logit mode or destination choice models that used inputs (e.g. skims) and created outputs (e.g. trip tables) in formats consistent with the modeling software. These were often stand-alone FORTRAN programs as some of the older modeling software was written in FORTRAN, which was considered some decades ago as the fastest programming language computationally. Many of these "legacy" programs are still in use in models around the U.S.

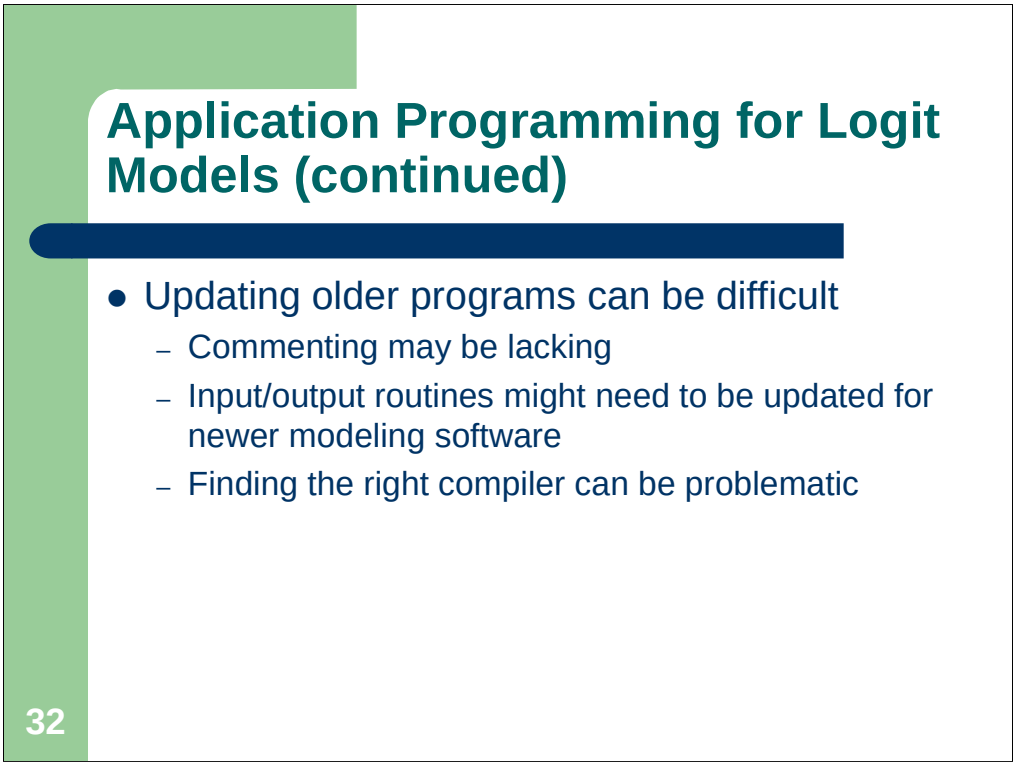


Key Message: Programming for Logit Models

Details:

There are many reasons why it is preferable to develop scripts using the scripting language in the modeling software. These include the following:

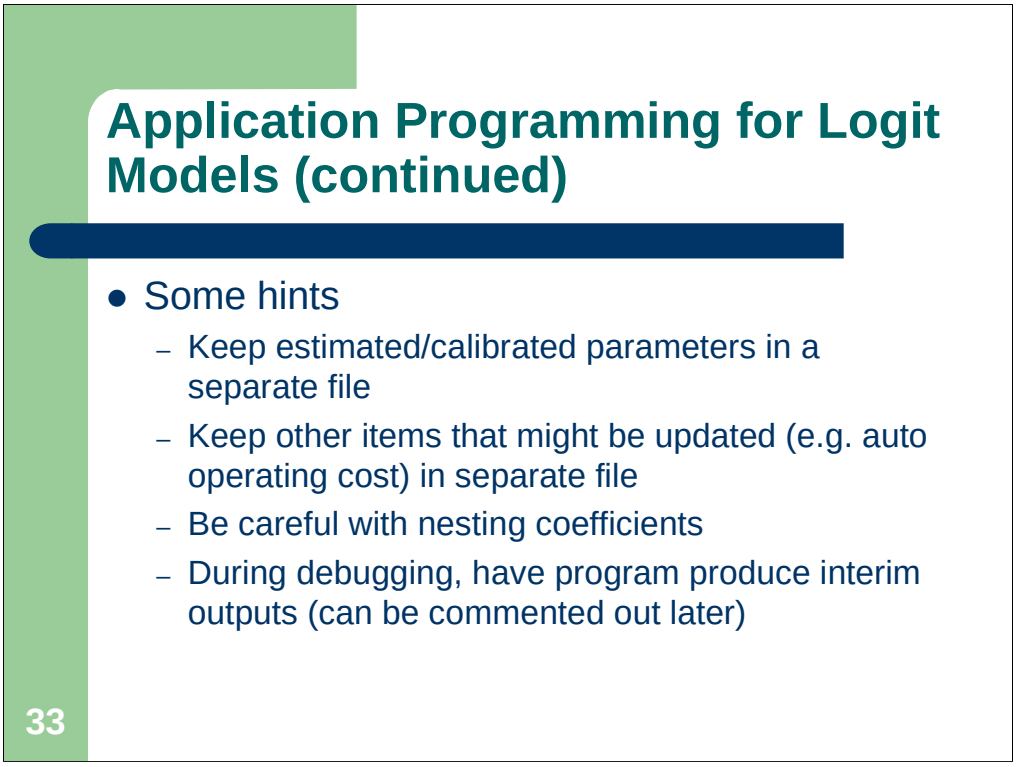
- It is easier to input/output skims, trip tables, etc. into the program smoothly
- It is easier to update the program when the model is updated (discussed further on next slide).
- Transparency - Many modelers are familiar with the modeling software, and so the programs are easier for others to understand.
- It can be easier to maintain quality control of the program and model results.
- One can get vendor support and assistance in creating/updating/maintaining the program.



Key Message: Programming for Logit Models

Details:

Updating older programs can be difficult. Commenting is often lacking. In many cases, the input/output routines for data such as skims might need to be updated to be consistent with the newer modeling software used by most agencies. Another issue is that there are different versions of languages such as FORTRAN, and recompiling the program after it has been changed may be problematic if the original compiler is no longer available.



Key Message: Programming for Logit Models

Details:

Here are some hints to improved scripting from our experience in implementing logit models:

Keep the estimated/calibrated parameters in a separate file, not hard coded into the model script. It is then clear where to go when the parameters are updated. It also saves recompiling when only the parameters are changing.

For similar reasons, keep other items that might be updated, such as auto operating costs, in a separate file.

Be careful with nesting coefficients. In application of a nested model, the utilities of lower level alternatives are multiplied by the nest coefficients.

During debugging, have the model program produce interim outputs that can be used to help with quality assurance and debugging. The statements that produce these interim outputs can always be commented out later.